

Gradient Support Projection Algorithm for Affine Feasibility Problem with Sparsity and Nonnegativity

Lili Pan^{1,2}, Naihua Xiu¹, Shenglong Zhou^{1*}

1. Department of Applied Mathematics Beijing Jiaotong University, Beijing 100044, P. R. China

2. Department of Mathematics, Shandong University of Technology, Zibo 255049, P. R. China

Abstract

Let A be a real $M \times N$ measurement matrix and $b \in \mathbb{R}^M$ be an observations vector. The affine feasibility problem with sparsity and nonnegativity (AFP_{SN} for short) is to find a sparse and non-negative vector $x \in \mathbb{R}^N$ with $Ax = b$ if such x exists. In this paper, we focus on establishment of optimization approach to solving the AFP_{SN} . By discussing tangent cone and normal cone of sparse constraint, we give the first necessary optimality conditions, α -Stability, T-Stability and N-Stability, and the second necessary and sufficient optimality conditions for the related minimization problems with the AFP_{SN} . By adopting Armijo-type stepsize rule, we present a framework of gradient support projection algorithm for the AFP_{SN} and prove its full convergence when matrix A is s -regular. By doing some numerical experiments, we show the excellent performance of the new algorithm for the AFP_{SN} without and with noise.

Keywords: affine feasibility problem; sparsity and nonnegativity; gradient support projection algorithm; s -regularity; numerical experiment

1 Introduction

In this paper, we mainly study an optimization approach to solving the affine feasibility problem with sparsity and nonnegativity (AFP_{SN}) defined by

$$\text{Find the vector } x \in \mathbb{R}^N \text{ with } x \geq 0, \|x\|_0 \leq s \text{ such that } Ax = b \quad (1)$$

if such x exists, where $A \in \mathbb{R}^{M \times N}$, $b \in \mathbb{R}^M$, $s < M < N$ and $\|x\|_0$ is the l_0 -norm of x , which refers to the number of non-zero elements in the vector x . Vector x is said to be s -sparse if $\|x\|_0 \leq s$. For $x = (x_1, \dots, x_N)^T, y = (y_1, \dots, y_N)^T \in \mathbb{R}^N$, $x \geq y$ stands for $x_i \geq y_i, i = 1, 2, \dots, N$.

*Corresponding author: Shenglong Zhou (longnan_zsl@163.com); Other two authors: Lili Pan (panlili1979@163.com), Naihua Xiu (nhxiu@bjtu.edu.cn). Time: June 27, 2014.

This is a class of inverse problems and has been popular for several years due to their applications in signal and image processing [8, 15], machine learning [17] and pattern recognition [5], and so on. For example, in many real-world problems the underlying parameters x represent quantities that can take on only nonnegative values, e.g., pixel intensities, frequency counts. In such cases, sparse affine feasibility problem must include nonnegative constraint on the model parameters x .

Usually, the AFP_{SN} is reformulated as the following optimization problem:

$$\begin{aligned} \min \quad & \frac{1}{2} \|Ax - b\|^2 \\ \text{s.t.} \quad & \|x\|_0 \leq s, x \geq 0. \end{aligned} \tag{2}$$

Let $S \triangleq \{x \in \mathbb{R}^N \mid \|x\|_0 \leq s\}$, then the feasible region of (2) is denoted as $S \cap \mathbb{R}_+^N$; here, $\|\cdot\|$ is l_2 -norm.

Greedy methods for (2) without nonnegativity have recently attracted much attention. One advantage of greedy methods is that they are generally faster than the relaxation approaches, and they can also be used to recover signals with more complex structures than sparsity such as tree sparse signals [3]. Another advantage of these methods is that many of them have stable recovery properties under certain conditions [11]. A variety of greedy methods have been proposed to tackle the so-called l_0 -problem, such as matching pursuit (MP) [18], orthogonal MP(OMP) [14], compressive sampling matching pursuit (CoSaMP) [19] and Subspace pursuit(SP) [13]. In [2], CoSaMP algorithm was extended to the objective function with arbitrary form. More recently, iterative hard thresholding algorithm (IHT) was proposed in [6]. Here, Beck et.al [4] showed that the limit points of the algorithm are L -stationary points if fixed stepsize $1/L$ is smaller than $\frac{1}{\lambda_{\max}(A^T A)}$. Blumensath [7] proposed an involved line-search method – normalised IHT (NIHT)– to adaptively select the stepsize per iteration. Cartis and Thompson [11] considered the convergence of IHT and NIHT from the aspect of recovery analysis [11]. Foucart [16] combined IHT and CoSaMP getting hard thresholding pursuit algorithm (HTP). While less effort has been made in sparsity and nonnegativity constraints simultaneously.

In this paper, we adopt a support projection method to solve this type of NP-hard problem starting from the iterative methods. Firstly, we study the tangent cone and normal cone of the sparse set under the Bouligand and Clarke concepts respectively. We propose three kinds of stability for sparsity constrained problems and analyze the relationship among them, which is α -Stability, T-Stability and N-Stability. We show that α -stability is most rigorous than the others. We also give the second order optimality condition for the same optimality problem under the concept of Clarke tangent cone. Secondly, we present a gradient support projection algorithm with Armijo-type's stepsize (GSPA) and prove the full convergent properties of the new algorithm under the condition that matrix A is s -regular. At last, numerical experiments demonstrate that GSPA performs very steadily whether for recovery without or with noise and is most time-saving compared with other three methods – NIHT, CoSaMP and SP.

This paper is organized as follows. Section 2 studies the first and second order optimality con-

ditions for a general sparse optimization model. Section 3 considers the corresponding results in Section 2 for sparsity and nonnegativity constrained problem (2). Section 4 gives the gradient support projection algorithm with Armijo-type stepsize and proves the convergence. Section 5 tests the performance of the new method. The last section gives some concluding remarks.

2 Optimality Conditions for Nonlinear Case

In this section, we study the first and second order optimality conditions for the following sparsity constrained nonlinear model:

$$\min f(x), \quad \text{s.t. } x \in S, \quad (3)$$

where $f(x) : \mathbb{R}^N \rightarrow \mathbb{R}$ is once or twice continuously differentiable.

We first consider the projection on sparse set S . For $S \subset \mathbb{R}^N$ being nonempty and closed, we call the mapping $P_S \Rightarrow S$ the projector onto S if

$$P_S(x) := \arg \min_{y \in S} \|x - y\|.$$

As S is nonconvex, the orthogonal projection operator $P_S(x)$ is not single-valued. It is well known that the sparse projection $P_S(x)$ sets all but s largest (in magnitude) elements of x to zero. If there is no unique such set, a set can be selected either randomly or according to some predefined ordering. We define $I_s(x) := \{j_1, j_2, \dots, j_s\} \subseteq \{1, 2, \dots, N\}$ of indices of x with $\min_{i \in I_s(x)} |x_i| \geq \max_{i \notin I_s(x)} |x_i|$. Then

$$P_S(x) = \left\{ y \in \mathbb{R}^N \mid y_i = \begin{cases} x_i, & i \in I_s(x), \\ 0, & i \notin I_s(x). \end{cases} \right\}$$

2.1 Tangent Cone and Normal Cone

Recalling that for any nonempty set $\Omega \subseteq \mathbb{R}^N$, its *Bouligand Tangent Cone* $T_\Omega^B(\bar{x})$, *Clarke Tangent Cone* $T_\Omega^C(\bar{x})$ and corresponding *Normal Cones* $N_\Omega^B(\bar{x})$ and $N_\Omega^C(\bar{x})$ at point $\bar{x} \in \Omega$ are defined as [20]:

$$\begin{aligned} T_\Omega^B(\bar{x}) &:= \left\{ d \in \mathbb{R}^N \mid \exists \{x^k\} \subset \Omega, \lim_{k \rightarrow \infty} x^k = \bar{x}, \lambda_k \geq 0, k = 1, 2, \dots \text{ such that } \lim_{k \rightarrow \infty} \lambda_k (x^k - \bar{x}) = d \right\}, \\ T_\Omega^C(\bar{x}) &:= \left\{ d \in \mathbb{R}^N \mid \begin{array}{l} \text{For } \forall \{x^k\} \subset \Omega, \forall \{\lambda_k\} \subset \mathbb{R}_+ \text{ with } \lim_{k \rightarrow \infty} x^k = \bar{x}, \lim_{k \rightarrow \infty} \lambda_k = 0, \\ \exists \{y^k\} \text{ such that } \lim_{k \rightarrow \infty} y^k = d \text{ and } x^k + \lambda_k y^k \in \Omega, k = 1, 2, \dots \end{array} \right\}, \\ N_\Omega^B(\bar{x}) &:= \{ d \in \mathbb{R}^N \mid \langle d, z \rangle \leq 0, \forall z \in T_\Omega^B(\bar{x}) \}, \\ N_\Omega^C(\bar{x}) &:= \{ d \in \mathbb{R}^N \mid \langle d, z \rangle \leq 0, \forall z \in T_\Omega^C(\bar{x}) \}. \end{aligned} \quad (4)$$

Theorem 2.1 For any $\bar{x} \in S$ and letting $\Gamma = \text{supp}(\bar{x})$, the Bouligand tangent cone and corresponding normal cone of S at \bar{x} are

$$T_S^B(\bar{x}) = \{ d \in \mathbb{R}^N \mid \|d\|_0 \leq s, \|\bar{x} + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R} \} \quad (5)$$

$$= \bigcup_Y \text{span}\{e_i, i \in Y \supseteq \Gamma, |Y| \leq s\} \quad (6)$$

$$N_S^B(\bar{x}) = \begin{cases} \{ d \in \mathbb{R}^N \mid d_i = 0, i \in \Gamma \} = \text{span}\{e_i, i \notin \Gamma\}, & \text{if } |\Gamma| = s \\ \{0\}, & \text{if } |\Gamma| < s \end{cases} \quad (7)$$

where $e_i \in \mathbb{R}^N$ is a vector whose the i th component is one and others are zeros, $\text{span}\{e_i, i \in \Gamma\}$ denotes the subspace of \mathbb{R}^N spanned by $\{e_i, i \in \Gamma\}$, and $\text{supp}(x) = \{i \in \{1, \dots, N\} \mid x_i \neq 0\}$.

Proof It is not difficult to verify that the right hand of (5) is equal to (6), and thus we only prove (5). First we prove $T_S^B(\bar{x}) \subseteq \{ d \in \mathbb{R}^N \mid \|d\|_0 \leq s, \|\bar{x} + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R} \}$. For $\forall d \in T_S^B(\bar{x})$, there is $\lim_{k \rightarrow \infty} x^k = \bar{x}$, $x^k \in S$, $\lambda_k \geq 0$ satisfies $d = \lim_{k \rightarrow \infty} \lambda_k(x^k - \bar{x})$. Since $\lim_{k \rightarrow \infty} x^k = \bar{x}$, there is k_0 when $k \geq k_0$, $\Gamma \subseteq \text{supp}(x^k)$. In addition, $d = \lim_{k \rightarrow \infty} \lambda_k(x^k - \bar{x})$ derives $\text{supp}(d) \subseteq \text{supp}(x^k)$, which combining with $\|x^k\|_0 \leq s$ when $k \geq k_0$ and $\Gamma \subseteq \text{supp}(x^k)$ yields that $\|d\|_0 \leq s$ and $\|\bar{x} + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R}$. Next we prove $T_S^B(\bar{x}) \supseteq \{ d \in \mathbb{R}^N \mid \|d\|_0 \leq s, \|\bar{x} + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R} \}$. For $\forall \|d\|_0 \leq s, \|\bar{x} + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R}$, we take any sequence $\{\lambda_k\}$ such that $\lambda_k > 0$ and $\lambda_k \rightarrow +\infty$. Then by defining $\{x^k\}$ with $x^k = \bar{x} + d/\lambda_k$, evidently $x^k \in S$, $\lim_{k \rightarrow \infty} x^k = \bar{x}$, and $d = \lim_{k \rightarrow \infty} \lambda_k(x^k - \bar{x})$, which implies $d \in T_S^B(\bar{x})$.

For (7), by the definition of $N_S^B(\bar{x})$, we obtain

$$\begin{aligned} N_S^B(\bar{x}) &= \{ d \in \mathbb{R}^N \mid \langle d, z \rangle \leq 0, \forall z \in T_S^B(\bar{x}) \} \\ &= \{ d \in \mathbb{R}^N \mid \langle d, z \rangle \leq 0, \|z\|_0 \leq s, \|\bar{x} + \mu z\|_0 \leq s, \forall \mu \in \mathbb{R} \}. \end{aligned} \quad (8)$$

If $|\Gamma| = s$, it yields $\text{supp}(z) \subseteq \Gamma$ for any $z \in T_S^B(\bar{x})$. Then

$$d \in N_S^B(\bar{x}) \iff \langle d, z \rangle \leq 0, \forall \text{supp}(z) \subseteq \Gamma \iff d_i \begin{cases} = 0, & i \in \Gamma, \\ \in \mathbb{R}, & i \notin \Gamma. \end{cases} \iff d \in \text{span}\{e_i, i \notin \Gamma\}.$$

If $|\Gamma| < s$, we will prove $N_S^B(\bar{x}) = \{0\}$. Assume $d \in N_S^B(\bar{x})$, we take $z = d_{i_0} e_{i_0}, \forall i_0 \in \{1, 2, \dots\}$, then $z \in T_S^B(\bar{x})$ since $|\Gamma| < s$. By $\langle d, z \rangle = d_{i_0}^2 \leq 0$, we can obtain $d_{i_0} = 0$. The arbitrariness of i_0 yields that $d = 0$, henceforth $N_S^B(\bar{x}) = \{0\}$. \square

Theorem 2.2 For any $\bar{x} \in S$ and letting $\Gamma = \text{supp}(\bar{x})$, then the Clarke tangent cone and corresponding normal cone of S at \bar{x} are

$$T_S^C(\bar{x}) = \{ d \in \mathbb{R}^N \mid \text{supp}(d) \subseteq \Gamma \} = \text{span}\{e_i, i \in \Gamma\} \quad (9)$$

$$N_S^C(\bar{x}) = \text{span}\{e_i, i \notin \Gamma\}. \quad (10)$$

Proof Obviously, $\text{span}\{e_i, i \in \Gamma\} = \{d \in \mathbb{R}^N \mid \text{supp}(d) \subseteq \Gamma\}$.

We first prove $T_S^C(\bar{x}) \subseteq \{d \in \mathbb{R}^N \mid \text{supp}(d) \subseteq \Gamma\}$. For $\forall d \in T_S^C(\bar{x})$, we have $\forall \{x^k\} \subset S, \forall \{\lambda_k\} \subset \mathbb{R}_+$ with $\lim_{k \rightarrow \infty} x^k = \bar{x}, \lim_{k \rightarrow \infty} \lambda_k = 0$, there is a sequence $\{y^k\}$ such that $\lim_{k \rightarrow \infty} y^k = d$ and $x^k + \lambda_k y^k \in S, k = 1, 2, \dots$. Assume that $\text{supp}(d) \not\subseteq \Gamma$, namely there is an $i_0 \in \text{supp}(d)$ but $i_0 \notin \Gamma$. Since $\lim_{k \rightarrow \infty} y^k = d$, it must have $y_{i_0}^k \rightarrow d_{i_0}$ which requires $y_{i_0}^k \neq 0$ when $k \geq k_0$. By the arbitrariness of $\{x^k\}$, we take $\{x^k\} \subset S$ such that $\lim_{k \rightarrow \infty} x^k = \bar{x}, \text{supp}(x^k) = \Gamma \cup \Gamma_k$ with $|\Gamma \cup \Gamma_k| = s$, where $\Gamma_k \subset \{1, 2, \dots, N\} \setminus \Gamma$, and $i_0 \notin \Gamma_k$. Because $\{y^k\}$ is fixed and the arbitrariness of $\{\lambda_k\}$, we can take $\{\lambda_k\}$ which satisfies $\lambda_k < 1$ and $\lambda_k / (\min_{i \in \Gamma \cup \Gamma_k, y_i^k \neq 0} |y_i^k|) \rightarrow 0$. Now we let

$$\lambda'_k = \min_{i \in \Gamma \cup \Gamma_k, y_i^k \neq 0} \lambda_k \left| \frac{x_i^k}{y_i^k} \right|.$$

Then $\lambda'_k (\neq 0) \rightarrow 0$ as $k \rightarrow \infty$. And thus $\forall i \in \Gamma \cup \Gamma_k$, either $|x_i^k + \lambda'_k y_i^k| = |x_i^k| > 0$ due to $y_i^k = 0$ or

$$|x_i^k + \lambda'_k y_i^k| \geq |x_i^k| - \lambda'_k |y_i^k| = |x_i^k| - |y_i^k| \min_{i \in \Gamma \cup \Gamma_k, y_i^k \neq 0} \lambda_k \left| \frac{x_i^k}{y_i^k} \right| \geq (1 - \lambda_k) |x_i^k| > 0.$$

Moreover, from $i_0 \notin \Gamma \cup \Gamma_k$ deriving $x_{i_0}^k = 0, y_{i_0}^k \neq 0$, we must have $\|x^k + \lambda'_k y^k\|_0 \geq s + 1$ for $k \geq k_0$, which is contradicted to $x^k + \lambda'_k y^k \in S$ for any $k = 1, 2, \dots$. Therefore $\text{supp}(d) \subseteq \Gamma$.

Next we prove $T_S^C(\bar{x}) \supseteq \{d \in \mathbb{R}^N \mid \text{supp}(d) \subseteq \Gamma\}$. For $\forall d \in \mathbb{R}^N$ such that $\text{supp}(d) \subseteq \Gamma$ and $\forall \{x^k\} \subset S, \forall \{\lambda_k\} \subset \mathbb{R}_+$ with $\lim_{k \rightarrow \infty} x^k = \bar{x}, \lim_{k \rightarrow \infty} \lambda_k = 0$, we have $\text{supp}(d) \subseteq \Gamma \subseteq \text{supp}(x^k)$ for any $k \geq k_0$. Let

$$\begin{aligned} y^k &= 0, & k &= 1, 2, \dots, k_0, \\ y^k &= x^k - \bar{x} + d, & k &= k_0 + 1, k_0 + 2, \dots, \end{aligned}$$

which brings out $x^k + \lambda_k y^k \in S$ for $k = 1, 2, \dots$ due to $x^k \in S$. In addition, $\lim_{k \rightarrow \infty} y^k = \lim_{t \rightarrow \infty} x^k - \bar{x} + d = d$. Hence $d \in T_S^C(\bar{x})$.

Finally (10) holding is obvious. Then the whole proof is completed. \square

Remark 2.1 Clearly for any $\bar{x} \in S$, Bouligand tangent cone $T_S^B(\bar{x})$ is closed but non-convex, while Clarke tangent cone $T_S^C(\bar{x})$ is closed and convex. In addition, $T_S^C(\bar{x}) \subseteq T_S^B(\bar{x})$.

2.2 α -Stability, N -Stability and T -Stability

When $f(x)$ is continuously differentiable on \mathbb{R}^N , we give the definition of three kinds of stability.

Definition 2.1 For real number $\alpha > 0$, a vector $x^* \in S$ is called an α -stationary point, N^\sharp -stationary point and T^\sharp -stationary point of (3) if it respectively satisfies the relation

$$\alpha\text{-stationary point: } x^* \in P_S(x^* - \alpha \nabla f(x^*)), \quad (11)$$

$$N^\sharp\text{-stationary point: } 0 \in \nabla f(x^*) + N_S^\sharp(x^*), \quad (12)$$

$$T^\sharp\text{-stationary point: } 0 = \|\nabla_S^\sharp f(x^*)\|, \quad (13)$$

where $\nabla_S^\# f(x^*) = \arg\min\{\|x + \nabla f(x^*)\| \mid x \in T_S^\#(x^*)\}$, $\# \in \{B, C\}$ stands for the sense of Bouligand tangent cone or Clarke tangent cone.

Theorem 2.3 Under the concept of Bouligand tangent cone, we consider model (3). For $\alpha > 0$, if the vector $x^* \in S$ satisfies $\|x^*\|_0 = s$, then

$$\alpha - \text{stationary point} \implies N^B - \text{stationary point} \iff T^B - \text{stationary point};$$

if the vector $x^* \in S$ satisfies $\|x^*\|_0 < s$, then

$$\alpha - \text{stationary point} \iff N^B - \text{stationary point} \iff T^B - \text{stationary point} \iff \nabla f(x^*) = 0.$$

Proof Denote $\Gamma = \text{supp}(x^*)$. If x^* is an α -stationary point of model (3), then from Lemma 2.2 in [4], it holds

$$x^* \in P_S(x^* - \alpha \nabla f(x^*)) \iff |(\nabla f(x^*))_i| \begin{cases} = 0, & i \in \Gamma \\ \leq \frac{1}{\alpha} M_s(|x^*|), & i \notin \Gamma, \end{cases} \quad (14)$$

for any $\alpha > 0$, where $M_s(|x^*|)$ is the s th largest element of $|x^*|$.

Case 1. First we consider the case $\|x^*\|_0 = s$. Under such circumstance, if x^* is an N^B -stationary point of model (3), then by (7) in Theorem 2.1, we have

$$-\nabla f(x^*) \in N_S^B(x^*) \iff (\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases} \quad (15)$$

Moreover, $\|x^*\|_0 = s$ produces

$$\begin{aligned} \nabla_S^B f(x^*) &= \arg\min\{\|d + \nabla f(x^*)\| \mid d \in T_S^B(x^*)\} \\ &= \arg\min\{\|d + \nabla f(x^*)\| \mid \|d\|_0 \leq s, \|x^* + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R}\} \\ &= \arg\min\{\|d + \nabla f(x^*)\| \mid \text{supp}(d) \subseteq \Gamma\}, \end{aligned} \quad (16)$$

where the third equality holds due to $\|x^*\|_0 = s$. (16) is equivalent to

$$(\nabla_S^B f(x^*))_i = \begin{cases} -(\nabla f(x^*))_i, & i \in \Gamma, \\ 0, & i \notin \Gamma. \end{cases}$$

Therefore, if x^* is an T^B -stationary point of model (3), then from above

$$\nabla_S^B f(x^*) = 0 \iff (\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases} \quad (17)$$

Henceforth, from (14), (15) and (17), one can easily check that when $\|x^*\|_0 = s$

$$\alpha - \text{stationary point} \implies N^B - \text{stationary point} \iff T^B - \text{stationary point}.$$

Case 2. Now we consider the case $\|x^*\|_0 < s$. Under such circumstance, $M_s(|x^*|) = 0$, and thus if x^* is an α -stationary point of model (3), then from (14), it holds

$$x^* \in P_S(x^* - \alpha \nabla f(x^*)) \iff \nabla f(x^*) = 0. \quad (18)$$

Then when $\|x^*\|_0 < s$, $N_S^B(x^*) = \{0\}$ from (7), which implies $\nabla f(x^*) = 0$. Therefore, if x^* is an N^B -stationary point of model (3), then

$$0 \in \nabla f(x^*) + N_S^B(x^*) \iff \nabla f(x^*) = 0. \quad (19)$$

Finally, we prove $\nabla f(x^*) = 0 \iff \nabla_S^B f(x^*) = 0$ when $\|x^*\|_0 < s$. On one hand, if $\nabla f(x^*) = 0$, then

$$\begin{aligned} \nabla_S^B f(x^*) &= \operatorname{argmin}\{ \|d + \nabla f(x^*)\| \mid \|d\|_0 \leq s, \|x^* + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R} \} \\ &= \operatorname{argmin}\{ \|d\| \mid \|d\|_0 \leq s, \|x^* + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R} \} = 0. \end{aligned}$$

On the other hand, if $\nabla_S^B f(x^*) = 0$, then

$$0 = \nabla_S^B f(x^*) = \operatorname{argmin}\{ \|d + \nabla f(x^*)\| \mid \|d\|_0 \leq s, \|x^* + \mu d\|_0 \leq s, \forall \mu \in \mathbb{R} \}$$

leads to $\|\nabla f(x^*)\| \leq \|d + \nabla f(x^*)\|$ for any $\|d\|_0 \leq s$, $\|x^* + \mu d\|_0 \leq s$, $\forall \mu \in \mathbb{R}$. Particular, for $\forall i_0 \in \{1, 2, \dots, N\}$, we take d with $\operatorname{supp}(d) = \{i_0\}$. Apparently, $\|x^* + \mu d\|_0 \leq s$, $\forall \mu \in \mathbb{R}$ owing to $\|x^*\|_0 < s$. Then by valuing $d_{i_0} = -(\nabla f(x^*))_{i_0}$ and $d_i = 0$, $i \neq i_0$, we immediately get $(\nabla f(x^*))_{i_0} = 0$ because of $\|\nabla f(x^*)\| \leq \|\nabla f(x^*) - (\nabla f(x^*))_{i_0}\|$. Then by the arbitrariness of i_0 , it holds $\nabla f(x^*) = 0$. Therefore, if x^* is an T^B -stationary point of model (3), then

$$\nabla_S^B f(x^*) = 0 \iff \nabla f(x^*) = 0. \quad (20)$$

Henceforth, from (18), (19) and (20), one can easily check that when $\|x^*\|_0 < s$

$$\alpha - \text{stationary point} \iff N^B - \text{stationary point} \iff T^B - \text{stationary point} \iff \nabla f(x^*) = 0.$$

Overall, the whole proof is finished. \square

Based on the proof of Theorem 2.3, we use the following table to illustrate the relationship among these three stationary points under the concept of Bouligand tangent cone.

Table 1: The relationship among these three kinds of stationary points.

	$\ x^*\ _0 = s$	$\ x^*\ _0 < s$
α - stationary point $x^* \in P_S(x^* - \alpha \nabla f(x^*)) \iff$	$ (\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \leq \frac{1}{\alpha} M_s(x^*), & i \notin \Gamma, \end{cases}$	$\nabla f(x^*) = 0$
N^B - stationary point $-\nabla f(x^*) \in N_S^B(x^*) \iff$	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases}$	$\nabla f(x^*) = 0$
T^B - stationary point $\nabla_S^B f(x^*) = 0 \iff$	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases}$	$\nabla f(x^*) = 0$

Theorem 2.4 Under the concept of Clarke tangent cone, we consider model (3). For $\alpha > 0$, if $x^* \in S$ then

$$\alpha - \text{stationary point} \implies N^C - \text{stationary point} \iff T^C - \text{stationary point}.$$

Proof Denote $\Gamma = \text{supp}(x^*)$. If x^* is an α -stationary point of model (3), for any $\alpha > 0$, we have (14)

If x^* is an N^C -stationary point of model (3), then by (10), we have

$$-\nabla f(x^*) \in N_S^C(x^*) \iff (\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases} \quad (21)$$

Moreover, by (9), it follows

$$\nabla_S^C f(x^*) = \operatorname{argmin}\{ \|d + \nabla f(x^*)\| \mid d \in T_S^C(x^*) \} = \operatorname{argmin}\{ \|d + \nabla f(x^*)\| \mid \operatorname{supp}(d) \subseteq \Gamma \},$$

which is equivalent to

$$(\nabla_S^C f(x^*))_i = \begin{cases} -(\nabla f(x^*))_i, & i \in \Gamma, \\ 0, & i \notin \Gamma. \end{cases}$$

Therefore, if x^* is an T^C -stationary point of model (3), then from above

$$\nabla_S^C f(x^*) = 0 \iff (\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases} \quad (22)$$

Henceforth, from (14), (21) and (22), one can easily check

$$\alpha - \text{stationary point} \implies N^C - \text{stationary point} \iff T^C - \text{stationary point}.$$

Overall, the whole proof is finished. \square

Based on the proof of Theorem 2.4, we use the following table to illustrate the relationship among these three stationary points under the concept of Clarke tangent cone.

Table 2: The relationship among these three kinds of stationary points.

	$\ x^*\ _0 = s$	$\ x^*\ _0 < s$
$\alpha - \text{stationary point}$ $x^* \in P_S(x^* - \alpha \nabla f(x^*)) \iff$	$ (\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \leq \frac{1}{\alpha} M_s(\ x^*\), & i \notin \Gamma, \end{cases}$	$\nabla f(x^*) = 0$
$N^C - \text{stationary point}$ $-\nabla f(x^*) \in N_S^C(x^*) \iff$	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases}$	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases}$
$T^C - \text{stationary point}$ $\nabla_S^C f(x^*) = 0 \iff$	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases}$	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma, \end{cases}$

2.3 Second Order Optimality Conditions

In this subsection, we study the second order necessary and sufficient optimality of model (3) if $f(x)$ is twice continuously differentiable on \mathbb{R}^N and satisfies the following assumption.

Assumption 2.1 *The gradient of the objective function $f(x)$ is Lipschitz with constant L_f over \mathbb{R}^N :*

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|, \quad \forall x, y \in \mathbb{R}^N. \quad (23)$$

Theorem 2.5 (Second Order Necessary Optimality) *If $x^* \in S$ is the optimal solution of (3), then for $0 < \alpha < \frac{1}{L_f}$, x^* is also the α -stationary point, moreover,*

$$d^\top \nabla^2 f(x^*) d \geq 0, \quad \forall d \in T_S^C(x^*). \quad (24)$$

where $\nabla^2 f(x^*)$ is the Hessian matrix of f at x^* .

Proof Since x^* is the optimal solution of (3), it must be an α -stationary point of model (3) for $0 < \alpha < \frac{1}{L_f}$ (Theorem 2.2 in [4]). By (14) and (9), one can easily verify that

$$d^\top \nabla f(x^*) = 0, \quad \forall d \in T_S^C(x^*).$$

Moreover, for any $\tau > 0$ and $d \in T_S^C(x^*)$, by the optimality of x^* and equality above, we have

$$\begin{aligned} 0 &\leq f(x^* + \tau d) - f(x^*) \\ &= f(x^*) + \tau d^\top \nabla f(x^*) + \frac{\tau^2}{2} d^\top \nabla^2 f(x^*) d + o(\|d\|^2) - f(x^*) \\ &= \frac{\tau^2}{2} d^\top \nabla^2 f(x^*) d + o(\|d\|^2), \end{aligned}$$

which implies that

$$d^\top \nabla^2 f(x^*) d \geq 0, \quad \forall d \in T_S^C(x^*).$$

The desired result is acquired. \square

Theorem 2.6 (Second Order Sufficient Optimality) *If $x^* \in S$ is an α -stationary point of (3) and satisfies*

$$d^\top \nabla^2 f(x^*) d > 0, \quad \forall d \in T_S^C(x^*), \quad (25)$$

then x^ is the strictly locally optimal solution of (3). Moreover, there is a $\gamma > 0$ and $\delta > 0$, when any $x \in B(x^*, \delta) \cap S$, it holds*

$$f(x) \geq f(x^*) + \gamma \|x - x^*\|^2. \quad (26)$$

Proof We only prove the second conclusion. From Table 2, one can easily check

$$d^\top \nabla f(x^*) = 0, \quad \forall d \in T_S^C(x^*), \quad (27)$$

By assuming the conclusion does not hold, there must be a sequence $\{x^k\}$ with

$$\lim_{k \rightarrow \infty} x^k = x^* \quad \text{and} \quad \text{supp}(x^k) = \text{supp}(x^*)$$

such that

$$f(x^k) - f(x^*) \leq \frac{1}{k} \|x^k - x^*\|^2. \quad (28)$$

Denote $d^k = \frac{x^k - x^*}{\|x^k - x^*\|}$. Due to $\|\frac{x^k - x^*}{\|x^k - x^*\|} = 1$, there exists a convergent subsequence, without loss of generality, assuming $d^k \rightarrow \bar{d}$. $d^k \in T_S^C(x^*)$ and $\bar{d} \in T_S^C(x^*)$ due to $\text{supp}(x^k) = \text{supp}(x^*)$, which means $d^{k^\top} \nabla f(x^*) = 0$ by (27). From (28), we have

$$\begin{aligned} \frac{1}{k} &\geq \frac{1}{\|x^k - x^*\|^2} \left(f(x^k) - f(x^*) \right) \\ &= \frac{1}{\|x^k - x^*\|^2} \left((x^k - x^*)^\top \nabla f(x^k) + \frac{1}{2} (x^k - x^*)^\top \nabla^2 f(x^*) (x^k - x^*) + o(\|x^k - x^*\|^2) \right) \\ &= d^{k^\top} \nabla^2 f(x^*) d^k + \frac{1}{\|x^k - x^*\|} d^{k^\top} \nabla f(x^*) + o(1) \\ &= d^{k^\top} \nabla^2 f(x^*) d^k + o(1). \end{aligned} \quad (29)$$

Then take the limit of both side of (29), we obtain

$$0 = \lim_{k \rightarrow \infty} \frac{1}{k} \geq \lim_{k \rightarrow \infty} \left(d^{k^\top} \nabla^2 f(x^*) d^k + o(1) \right) = \bar{d}^\top \nabla^2 f(x^*) \bar{d} > 0, \quad \bar{d} \in T_S^C(x^*),$$

which is contradicted. Therefore the conclusion does hold. \square

3 Optimality Conditions for Model (2)

In this section, we mainly aim at specifying the results in Section 2 to the model (2). For notational simplicity, we hereafter denote $r(x) \triangleq \frac{1}{2} \|Ax - b\|^2$. First, we define the projection on $S \cap \mathbb{R}_+^N$ named **support projection**, which has an explicit expression.

Proposition 3.1 $P_{S \cap \mathbb{R}_+^N}(x) = P_S \cdot P_{\mathbb{R}_+^N}(x)$.

Proof Denote $I_+(x) = \{i | x_i > 0\}$, $I_0(x) = \{i | x_i = 0\}$, $I_-(x) = \{i | x_i < 0\}$, let $y \in P_{S \cap \mathbb{R}_+^N}(x)$. For $i \in I_0(x) \cup I_-(x)$, it is easy to see $y_i = 0$. There are two cases:

Case 1, $|I_+(x)| \leq s$, then $y = P_{\mathbb{R}_+^N}(x) = P_S \cdot P_{\mathbb{R}_+^N}(x)$.

Case 2, $|I_+(x)| > s$, we should choose no more than s coordinates from $I_+(x)$ to minimize $\|x - y\|$. For $i, j \in I_+(x)$ and $x_i > x_j$,

$$(x_i - x_i)^2 + (x_j - x_j)^2 < (x_i - x_i)^2 + (x_j - 0)^2 < (x_i - 0)^2 + (x_j - x_j)^2 < (x_i - 0)^2 + (x_j - 0)^2.$$

Then the projection on $S \cap \mathbb{R}_+^N$ sets all but s largest elements of $P_{\mathbb{R}_+^N}(x)$ to zero, which is $P_S \cdot P_{\mathbb{R}_+^N}(x)$. \square

Notice that the order of projections can't be changed. For example $x = (-2, 1)^T$, $s = 1$. $P_{S \cap \mathbb{R}_+^2}(x) = P_S \cdot P_{\mathbb{R}_+^2}(x) = (0, 1)^T$, while $P_{\mathbb{R}_+^2} \cdot P_S(x) = (0, 0)^T$.

The direct result of Theorem 2.1 and 2.2 is the following theorem.

Theorem 3.1 For any $\bar{x} \in S \cap \mathbb{R}_+^N$, by denoting $\mathbb{R}_+^N(\bar{x}) := \{x \in \mathbb{R}^N \mid x_i \geq 0, i \notin \Gamma\}$, it follows

$$T_{S \cap \mathbb{R}_+^N}^B(\bar{x}) = T_S^B(\bar{x}) \cap \mathbb{R}_+^N(\bar{x}), \quad N_{S \cap \mathbb{R}_+^N}^B(\bar{x}) = T_S^B(\bar{x}) \cap (-\mathbb{R}_+^N(\bar{x})) \quad (30)$$

$$T_{S \cap \mathbb{R}_+^N}^C(\bar{x}) = T_S^C(\bar{x}), \quad N_{S \cap \mathbb{R}_+^N}^C(\bar{x}) = N_S^C(\bar{x}). \quad (31)$$

For model (2), we have the corresponding definition of α -stationary point, N^\sharp -stationary point and T^\sharp -stationary point by substituting $S \cap \mathbb{R}_+^N$ for S , where $\sharp \in \{B, C\}$ stands for the sense of Bouligand tangent cone or Clarke tangent cone. In order to facilitate the discussion next, we describe a more explicit representation of α -stationary point, that is

$$x^* \in P_{S \cap \mathbb{R}_+^N}(x^* - \alpha \nabla r(x^*)). \quad (32)$$

Theorem 3.2 For any $\alpha > 0$, a vector $x^* \in S \cap \mathbb{R}_+^N$ is α -stationary point of (2) if and only if

$$\nabla_i r(x^*) \begin{cases} = 0, & \text{if } i \in \text{supp}(x^*), \\ \geq 0, \text{ or } \in [-\frac{1}{\alpha} M_s(x^*), 0], & \text{if } i \notin \text{supp}(x^*), \end{cases} \quad (33)$$

Proof Suppose (32) is satisfied for x^* . If $i \in \text{supp}(x^*)$, then $x_i^* = x_i^* - \alpha \nabla_i r(x^*)$, so that $\nabla_i r(x^*) = 0$; If $i \notin \text{supp}(x^*)$, there are two cases: either $x_i^* - \alpha \nabla_i r(x^*) \leq 0$, that is $\nabla_i r(x^*) \geq x_i^* = 0$, or $0 \leq x_i^* - \alpha \nabla_i r(x^*) \leq M_s(x^*)$, that is $-\frac{1}{\alpha} M_s(x^*) \leq \nabla_i r(x^*) \leq 0$.

On the contrary, assume (33) holds. If $\|x^*\|_0 < s$, we get $M_s(x^*) = 0$, then for $i \in \text{supp}(x^*)$, $\nabla_i r(x^*) = 0$, then $x^* - \alpha \nabla_i r(x^*) = x_i^*$ or for $i \notin \text{supp}(x^*)$, $x_i^* - \alpha \nabla_i r(x^*) \leq 0$, therefore, (32) holds. If $\|x^*\|_0 = s$, that is $M_s(x^*) > 0$. By (33), for $i \in \text{supp}(x^*)$, $\alpha \nabla_i r(x^*) = 0$; for $i \notin \text{supp}(x^*)$, $x^* - \alpha \nabla_i r(x^*) \leq 0$ or $0 \leq x_i^* - \alpha \nabla_i r(x^*) \leq M_s(x^*)$, so that (32) holds as well. \square

One can easily check that when $\|\bar{x}\|_0 = s$ it holds $T_{S \cap \mathbb{R}_+^N}^B(\bar{x}) = T_S^B(\bar{x})$. Therefore the corresponding theorem is derived by Theorems 2.3, 2.4, 3.2 and Corollary 3.1 directly.

Theorem 3.3 For the model (2) and any $\alpha > 0$.

A) Under the concept of Bouligand tangent cone, if $\|x^*\|_0 = s, x^* \geq 0$, then

$$\alpha - \text{stationary point} \implies N^B - \text{stationary point} \iff T^B - \text{stationary point}.$$

B) Under the concept of Clarke tangent cone, if $\|x^*\|_0 \leq s, x^* \geq 0$, then

$$\alpha - \text{stationary point} \implies N^C - \text{stationary point} \iff T^C - \text{stationary point}.$$

Combining Theorems 2.5 and 2.6, we derive the following second order optimality result.

Theorem 3.4 (Second Order Optimality) *If $x^* \in S \cap \mathbb{R}_+^N$ is the optimal solution of (2), then for $0 < \alpha < \frac{1}{\lambda_{\max}(A^T A)}$, x^* is also the α -stationary point of (2), and moreover,*

$$d^T A^T A d \geq 0, \quad \forall d \in T_S^C(x^*). \quad (34)$$

On the contrary, if $x^ \in S \cap \mathbb{R}_+^N$ is an α -stationary point of (2) and satisfies*

$$d^T A^T A d > 0, \quad \forall d \in T_S^C(x^*), \quad (35)$$

then x^ is the strictly locally optimal solution of (2). Moreover, there is a $\gamma > 0$ and $\delta > 0$, when any $x \in B(x^*, \delta) \cap S \cap \mathbb{R}_+^N$, it holds*

$$r(x) \geq r(x^*) + \gamma \|x - x^*\|^2. \quad (36)$$

4 Gradient Support Projection Algorithm

We now develop the gradient support projection algorithm with Armijo-type stepsize rule which is shortly denoted as GSPA. For simplicity, we utilize $L_r := \lambda_{\max}(A^T A)$ to denote the Lipschitz constant of $\nabla r(x)$.

Table 3: The framework of GSPA.

Step 0 Initialize $x^0 = 0$, $\Gamma^0 = \text{supp}(P_{S \cap \mathbb{R}_+^N}(A^T b))$, $0 < \alpha_0 < \frac{1}{L_r}$, $0 < \sigma \leq \frac{1}{4L_r}$, $0 < \beta < 1$, $\epsilon > 0$. Set $k \leftarrow 0$;

Step 1 Compute $\tilde{x}^{k+1} = P_{S \cap \mathbb{R}_+^N}(x^k - \alpha_0 \nabla r(x^k))$;

Step 2 If $\text{supp}(\tilde{x}^{k+1}) = \Gamma^k$, then $x^{k+1} = \tilde{x}^{k+1}$, $\Gamma^{k+1} = \text{supp}(x^{k+1})$;

Else compute $x^{k+1} = P_{S \cap \mathbb{R}_+^N}(x^k - \alpha_k \nabla r(x^k))$, $\Gamma^{k+1} = \text{supp}(x^{k+1})$, where $\alpha_k = \alpha_0 \beta^{m_k}$ and m_k is the smallest positive integer m such that

$$r(x^k(\alpha_0 \beta^m)) \leq r(x^k) - \frac{\sigma}{2} \frac{\|x^k(\alpha_0 \beta^m) - x^k\|^2}{(\alpha_0 \beta^m)^2},$$

here $x^k(\alpha) = P_{S \cap \mathbb{R}_+^N}(x^k - \alpha \nabla r(x^k))$;

Step 3 If $\|x^{k+1} - x^k\| \leq \epsilon$, then stop; Otherwise $k \leftarrow k + 1$, go to **Step 1**.

Remark Compared with IHT in [4], we mainly add Armijo-type stepsize rule in Step 2, which is well defined by Lemma 4.1.

Lemma 4.1 *Let $\{x^k\}$ be the iterative point in Step 2 in GSPA. Then*

$$r(x^k(\alpha)) \leq \begin{cases} r(x^k) - \frac{1}{2}(\frac{1}{\alpha} - L_r) \|x^k(\alpha) - x^k\|^2, & \alpha \in (0, \frac{1}{L_r}) \\ r(x^k) - \frac{\sigma}{2} \frac{\|x^k(\alpha) - x^k\|^2}{\alpha^2}, & \alpha \in \left[\frac{1 - \sqrt{1 - 4\sigma L_r}}{2L_r}, \frac{1 + \sqrt{1 - 4\sigma L_r}}{2L_r} \right]. \end{cases} \quad (37)$$

Proof From the algorithm in Step 2, we have

$$x^k(\alpha) \in \operatorname{argmin} \left\{ \|x - x^k + \alpha \nabla r(x^k)\|^2, \|x\|_0 \leq s, x \geq 0 \right\},$$

which implies that $\|x^k(\alpha) - x^k + \alpha \nabla r(x^k)\|^2 \leq \|\alpha \nabla r(x^k)\|^2$, that is

$$\|x^k(\alpha) - x^k\|^2 \leq -2\alpha \langle \nabla r(x^k), x^k(\alpha) - x^k \rangle. \quad (38)$$

From

$$\begin{aligned} r(x^k(\alpha)) &= r(x^k) + \langle \nabla r(x^k), x^k(\alpha) - x^k \rangle + \frac{1}{2} \|A(x^k(\alpha) - x^k)\|^2 \\ &\leq r(x^k) - \frac{1}{2\alpha} \|x^k(\alpha) - x^k\|^2 + \frac{Lr}{2} \|x^k(\alpha) - x^k\|^2 \end{aligned} \quad (39)$$

we can obtain the desired result by the definition of α . \square

Using Lemma 4.1 and other properties of iterative sequence, the convergence properties of GSPA can be established.

Theorem 4.1 *Let the sequence $\{x^k\}$ be generated by GSPA, we have*

- (i) $\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^k\|}{\alpha_k} = 0$;
- (ii) *any accumulation point of $\{x^k\}$ is the α -stationary point of (2);*
- (iii) $\lim_{k \rightarrow \infty} \|\nabla_{S \cap \mathbb{R}_+^N}^C r(x^k)\| = 0$.

Proof (i) From (37), we derive that $r(x^k) - r(x^{k+1}) \geq c \frac{\|x^{k+1} - x^k\|^2}{\alpha_k^2}$, where $c = \min\{\frac{\alpha_0 - Lr\alpha_0^2}{2}, \frac{\alpha}{2}\}$. Then

$$\sum_{k=0}^{\infty} \frac{\|x^{k+1} - x^k\|^2}{\alpha_k^2} \leq \frac{1}{c} \sum_{k=0}^{\infty} (r(x^k) - r(x^{k+1})) = \frac{1}{c} r(x^0) < +\infty,$$

which signifies $\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^k\|}{\alpha_k} = 0$. Since α_k is bounded from below by a positive constant, we conclude that $\lim_{k \rightarrow \infty} \|x^{k+1} - x^k\| = 0$.

(ii) Suppose that x^* is an accumulate point of the sequence $\{x^k\}$, then there exists a subsequence $\{x^{k_n}\}$ that converges to x^* . By (i), $\lim_{n \rightarrow \infty} x^{k_n+1} = x^*$. Based on $x^{k_n+1} = P_{S \cap \mathbb{R}_+^N}(x^{k_n} - \alpha \nabla r(x^{k_n}))$ in Step 2, we consider two cases.

Case 1. $i \in \operatorname{supp}(x^*)$. The convergence of $\{x^{k_n}\}$ and $\{x^{k_n+1}\}$ guarantees that for some $n_1 > 0$, $x_i^{k_n} > 0$, $x_i^{k_n+1} > 0$ for all $n > n_1$. The definition of the projection on $S \cap \mathbb{R}_+^N$ shows that

$$x_i^{k_n+1} = x_i^{k_n} - \alpha \nabla_i r(x^{k_n}).$$

Taking $n \rightarrow \infty$, we have $\nabla_i r(x^*) = 0$.

Case 2. $i \notin \operatorname{supp}(x^*)$. If there exists an $n_2 > 0$ such that for all $n > n_2$, $x_i^{k_n+1} = 0$, the projection implies that

$$x_i^{k_n} - \alpha \nabla_i r(x^{k_n}) \leq 0 \text{ or } 0 < x_i^{k_n} - \alpha \nabla_i r(x^{k_n}) \leq M_s(x^{k_n+1}).$$

Letting $n \rightarrow \infty$ and exploiting the continuity of the function M_s , we obtain that

$$\nabla_i r(x^*) \geq 0 \text{ or } -M_s(x^*) \leq \alpha \nabla_i r(x^*) \leq 0.$$

On the other hand, if there exists an infinite number of indices of k_n for $x_i^{k_n+1} > 0$, as the same proof in Case 1, it follows that $\nabla_i r(x^*) = 0$. Since α_k is bounded from below by a positive constant, we have

$$\nabla_i r(x^*) \begin{cases} = 0, & \text{if } i \in \text{supp}(x^*), \\ \geq 0, \text{ or } \in [-\frac{1}{\alpha} M_s(x^*), 0], & \text{if } i \notin \text{supp}(x^*); \end{cases} \quad (40)$$

which means x^* is an α -stationary point of (2) by Theorem 3.2.

(iii) From Theorem 3.1, $T_{S \cap \mathbb{R}_+^N}^C(x^k) = \{d \in \mathbb{R}^N \mid \text{supp}(d) \subseteq \text{supp}(x^k)\}$ is a subspace. Then

$$\|\nabla_{S \cap \mathbb{R}_+^N}^C r(x^k)\| = \max\{\langle -\nabla r(x^k), v^k \rangle \mid v^k \in T_{S \cap \mathbb{R}_+^N}^C(x^k), \|v^k\| \leq 1\},$$

see Lemma 3.1 in [10]. we have for any $\eta > 0$, there is a $v^k \in T_{S \cap \mathbb{R}_+^N}^C(x^k)$ with $\|v^k\| = 1$ satisfying

$$\|\nabla_{S \cap \mathbb{R}_+^N}^C r(x^k)\| \leq -\langle \nabla r(x^k), v^k \rangle + \eta. \quad (41)$$

For all $z^{k+1} \in T_{S \cap \mathbb{R}_+^N}^C(x^{k+1})$ and $x^{k+1} = P_{S \cap \mathbb{R}_+^N}(x^k - \alpha_k \nabla r(x^k))$, $x^{k+1} - (x^k - \alpha_k \nabla r(x^k))$ is orthogonal to $T_{S \cap \mathbb{R}_+^N}^C(x^{k+1})$, which yields that

$$\langle x^{k+1} - (x^k - \alpha_k \nabla r(x^k)), z^{k+1} - x^{k+1} \rangle = 0.$$

Letting $v^{k+1} = \frac{z^{k+1} - x^{k+1}}{\|z^{k+1} - x^{k+1}\|} \in T_{S \cap \mathbb{R}_+^N}^C(x^{k+1})$, with Cauchy-Schwartz inequality, the above equation leads to

$$-\langle \nabla r(x^k), v^{k+1} \rangle \leq \frac{\|x^{k+1} - x^k\|}{\alpha_k}. \quad (42)$$

From (i), $\limsup_{k \rightarrow \infty} -\langle \nabla r(x^k), v^{k+1} \rangle \leq 0$. Combining

$$-\langle \nabla r(x^{k+1}), v^{k+1} \rangle = -\langle \nabla r(x^{k+1}) - \nabla r(x^k), v^{k+1} \rangle - \langle \nabla r(x^k), v^{k+1} \rangle$$

with (42), (i) and Lipschitz continuity of $\nabla r(x)$, we have

$$\limsup_{k \rightarrow \infty} -\langle \nabla r(x^{k+1}), v^{k+1} \rangle \leq 0.$$

By (41) and the arbitrariness of η , we can prove the result. \square

In order to attain the result that $\{x^k\}$ converges to a local minimizer of (2), we need the following assumption and lemma.

Assumption 4.1 ([4]) *Matrix A is s -regular if any s of its columns are linearly independent, namely,*

$$d^\top A^\top A d > 0, \quad \forall \|d\|_0 \leq s. \quad (43)$$

Theorem 4.2 *Let the sequence $\{x^k\}$ be generated by GSPA, then $\{x^k\}$ converges to a local minimizer of (2) if Assumption 4.1 holds.*

Proof For completeness, we give the proof similar to that of Theorem 3.2 in [4].

First, the number of α -stationary points of (2) is finite. In fact, by Theorem 3.2, α -stationary point x^* satisfies

$$\nabla_{\Gamma} r(x^*) = A_{\Gamma}^T (A_{\Gamma} x_{\Gamma} - b) = 0, \Gamma = \text{supp}(x^*), x_{\Gamma} \geq 0, |\Gamma| \leq s,$$

which has at most one solution. Since the number of subsets of $\{1, 2, \dots, N\}$ whose size is no larger than s is $T = \sum_{i=0}^s C_N^i$, the number of α -stationary points of (2) is no more than T .

Now we show $\{x^k\}$ is bounded. Lemma 4.1 indicates that $\{r(x^k)\}$ is decreasing, then the sequence $\{x^k\}$ is contained in the level set

$$E = \{x \in \mathbb{R}_+^N \cap S \mid r(x) \leq r(x^0)\}.$$

We can represent the set E as the union $E = \bigcup_{j=1}^T E_j$, where $E_j = \{x \in \mathbb{R}_+^N \mid \|Ax - b\|^2 \leq r(x^0), x_i = 0, i \notin \Gamma_j, |\Gamma_j| \leq s\}$. Since Assumption 4.1 implies that $A_{\Gamma_j}^T A_{\Gamma_j}$ is positive definite, E_j is bounded, which shows E is bounded. Combining the boundedness of $\{x^k\}$ and (ii) in Theorem 4.1, we obtain that there is a subsequence $\{x^{k_n}\}$ converges to an α -stationary point x^* .

We conclude that $\lim_{k \rightarrow \infty} x^k = x^*$. Since the number of α -stationary points of (2) is finite, there exists an $\epsilon > 0$ smaller than the minimal distance between all the pairs of the α -stationary points. We show the convergence of x^k by contradiction. When n is sufficiently large, $\|x^{k_n} - x^*\| \leq \epsilon$, without loss of generality, we assume the above inequality holds for all $n \geq 0$. Since x^k is divergent, the index l_n given by

$$l_n = \min\{i \mid \|x^i - x^*\| > \epsilon, i > k_n, i \in \mathbb{N}\}$$

is well defined. We have thus constructed a subsequence $\{x^{l_n}\}$ for which

$$\|x^{l_{n-1}} - x^*\| \leq \epsilon, \|x^{l_n} - x^*\| > \epsilon, n = 1, 2, \dots$$

It follows that $\{x^{l_{n-1}}\}$ converges to x^* , there exists an $n_0 > 0$ such that for all $n > n_0$, $\|x^{l_{n-1}} - x^*\| \leq \frac{\epsilon}{2}$. Then for all $n > n_0$, $\|x^{l_{n-1}} - x^{l_n}\| > \frac{\epsilon}{2}$, contradicting (i) in Theorem 4.1.

Finally, by s -regularity of A and Theorem 3.4, it has that x^* is also the local minimizer of (2). Therefore $\{x^k\}$ converges to a local minimizer of (2). \square

By analyzing the convergence theorems, we can obtain the theorem of existence of optimal solution of (2), which can be regarded as the theorem of second order sufficient optimality condition.

Theorem 4.3 *Suppose that Assumption 4.1 holds for matrix A , then the local solutions of problem (2) exist and are finite. Moreover, its global solution exists consequently.*

Remark We achieve the stronger convergence results (Theorem 4.1, 4.2 and 4.3) under relatively weaker assumption compared with [1, 7]. More exactly, the gradient projection algorithm in [1] converges to a N -stationary point provided the iteration sequence is bounded, while GSPA has the same result without boundedness of the iterative sequence. NIHT in [7] converges to a local minimizer of (2) if A satisfies the restricted isometry property (RIP) (introduced in [9]), while GSPA has the same convergence result with s -regularity of A , which is weaker than RIP.

5 Numerical Experiments

Before proceeding to the computational results, we need to define some notations and data sets. For convenience and clear understanding in the graph presentations and some comments, we use the notations: *GSPA*, *NIHT*, *CSMP* (short for *CoSaMP*) and *SP* to represent the our Gradient Support Projection Algorithm, Normalized Iterative Hard Thresholding (proposed by Blumensath in [7]), Compressive Sampling Matching Pursuit (established by Thomas et al. in [19]), and Subspace Pursuit in [13] respectively without nonnegative constraints. We first will compare the numerical performance of *GSPA* and *NIHT* under the nonnegative constraints, and thus denote them as N_GSPA and N_NIHT .

To accelerate the rate of convergence, α_0 in each iteration is chosen according to [7]

$$\alpha_0^k = \frac{\|A_{\Gamma^k}^T(b - Ax^k)\|^2}{\|A_{\Gamma^k} A_{\Gamma^k}^T(b - Ax^k)\|^2},$$

For each data set, the random matrix A and the designed vector b in absence of nonnegative constraints are generated by the following MATLAB codes:

$$\begin{aligned} x_{\text{orig}} &= \text{zeros}(N, 1), \quad y = \text{randperm}(N), \quad x_{\text{orig}}(y(1:s)) = \text{randn}(s, 1), \\ A &= \text{randn}(M, N), \quad b = A * x_{\text{orig}}. \end{aligned}$$

If considering the nonnegative constraints, we simply alter corresponding code as

$$x_{\text{orig}}(y(1:s)) = \text{abs}(\text{randn}(s, 1)).$$

where the sparsity s is taking $s = 1\%N$ or $k = 5\%N$. In terms of parameters, we fix $\beta = 0.8$ and $\sigma = 10^{-5}$ for simplicity. For each data set, we will randomly run 40 samples and the stopping criterias will be set by $\|x^{k+1} - x^k\| \leq 10^{-6}$ or the maximum iterative times is equal to 5000 for all methods. In the following analysis, we say x as the recovered solution from the affine equations. In whole experiments, the average prediction error $\|Ax - b\|$, the recovered error $\|x - x_{\text{orig}}\|_{\infty}$ and CPU time will be taken into consideration to illustrate the performance of the four methods. All those simulations are carried out on a CPU 2.6GHz laptop.

5.1 Comparison of N_NIHT and N_GSPA

We first will compare N_NIHT (which can be simply altered by adding the nonnegative constraints $x \geq 0$ when pursuing the projection in $NIHT$) and N_GSPA by setting different N with $s = 5\%N$ and running 40 samples for each data set.

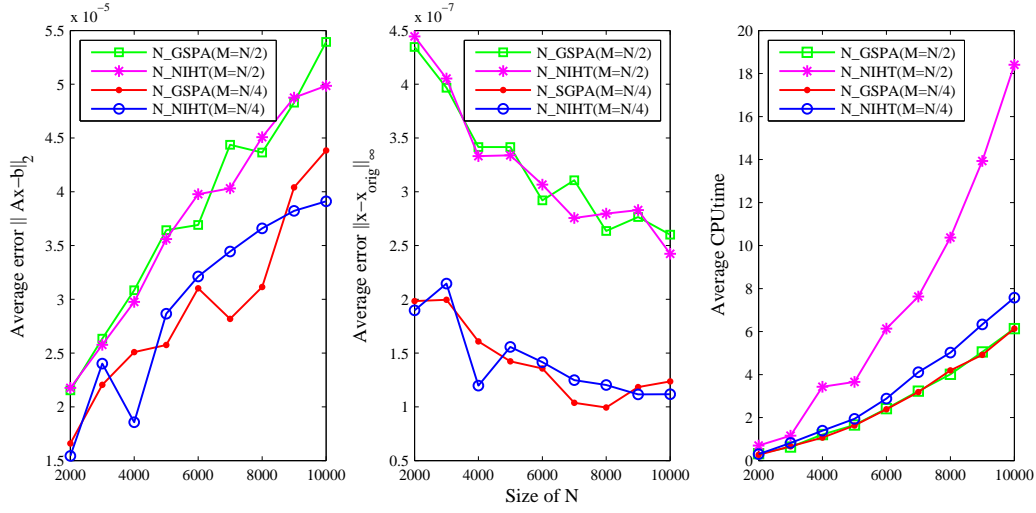


Figure 1: Average results yielded by N_NIHT and N_GSPA under $M = N/2$ and $M = N/4$.

The corresponding results can be seen in Figure 1, from which one can discern that when $M = N/2$ and $M = N/4$ there is no big difference of these two methods. To be more exact, the average prediction error $\|Ax-b\|$ and recovered error $\|x-x_{orig}\|_{\infty}$ are quite small with order of magnitude 10^{-5} and 10^{-7} , which means the recover is almost exact. In terms of the average CPU time, see the third column of Figure 1. For one thing, the time cost by N_GSPA is smaller than that from N_NIHT for each case. For another, one can check that N_NIHT are relatively dependent on the sample dimension M . The larger the M is, the smaller is the time, which implies the performance of N_NIHT tends to be worse with the M decreasing for fixed N . By contrast, the performance of our N_GSPA approach is much more stable since the time cost by this method is nearly similar under different M .

5.2 Comparison of $GSPA$, $NIHT$, $CSMP$ and SP

In the sequent part, we mainly compare $GSPA$, $NIHT$, $CSMP$ and SP without the nonnegative constraints. The primal MATLAB codes of $CSMP$ and SP can be download from the website below:

http://media.aau.dk/null_space_pursuits/2011/07/a-few-corrections-to-cosamp-and-sp-matlab.html.

Exact Recovery: We firstly consider the exact sparse recovery $b = Ax_{orig}$. Through running 40 examples, the produced data is listed as Tables 4–6.

Table 4: The average prediction error $\|Ax - b\|$ over 40 simulations with $s = 5\%N$.

N	M	$GSPA$	$NIHT$	$CSMP$	SP
$N = 1000$	$M = N/4$	0.14e-04	0.15e-04	0.00e-04	0.00e-04
	$M = N/2$	0.09e-04	0.09e-04	0.00e-04	0.00e-04
$N = 3000$	$M = N/4$	0.29e-04	0.27e-04	0.00e-04	0.00e-04
	$M = N/2$	0.17e-04	0.20e-04	0.00e-04	0.00e-04
$N = 5000$	$M = N/4$	0.36e-04	0.34e-04	0.00e-04	0.00e-04
	$M = N/2$	0.23e-04	0.24e-04	0.00e-04	0.00e-04
$N = 7000$	$M = N/4$	0.42e-04	0.41e-04	0.00e-04	0.00e-04
	$M = N/2$	0.29e-04	0.32e-04	0.00e-04	0.00e-04
$N = 10000$	$M = N/4$	0.51e-04	0.48e-04	0.00e-04	0.00e-04
	$M = N/2$	0.37e-04	0.39e-04	0.00e-04	0.00e-04

Table 5: The average recovered error $\|x_{\text{orig}} - x\|_{\infty}$ over 40 simulations with $s = 5\%N$.

N	M	$GSPA$	$NIHT$	$CSMP$	SP
$N = 1000$	$M = N/4$	0.50e-06	0.51e-06	0.00e-06	0.00e-06
	$M = N/2$	0.20e-06	0.16e-06	0.00e-06	0.00e-06
$N = 3000$	$M = N/4$	0.45e-06	0.38e-06	0.00e-06	0.00e-06
	$M = N/2$	0.16e-06	0.17e-06	0.00e-06	0.00e-06
$N = 5000$	$M = N/4$	0.31e-06	0.30e-06	0.00e-06	0.00e-06
	$M = N/2$	0.12e-06	0.11e-06	0.00e-06	0.00e-06
$N = 7000$	$M = N/4$	0.26e-06	0.26e-06	0.00e-06	0.00e-06
	$M = N/2$	0.12e-06	0.12e-06	0.00e-06	0.00e-06
$N = 10000$	$M = N/4$	0.24e-06	0.24e-06	0.00e-06	0.00e-06
	$M = N/2$	0.10e-06	0.10e-06	0.00e-06	0.00e-06

Table 6: The average CPU time over 40 simulations with $s = 5\%N$.

N	M	$GSPA$	$NIHT$	$CSMP$	SP
$N = 1000$	$M = N/4$	0.0689	0.2583	0.1492	0.0961
	$M = N/2$	0.0677	0.2459	0.1687	0.1307
$N = 3000$	$M = N/4$	0.5385	3.3210	1.9171	1.1197
	$M = N/2$	0.5756	2.6228	1.8754	1.3627
$N = 5000$	$M = N/4$	1.5583	11.246	8.0507	4.5900
	$M = N/2$	1.5114	8.0690	7.7457	5.0981
$N = 7000$	$M = N/4$	3.0050	20.761	19.698	10.729
	$M = N/2$	2.9543	16.389	19.336	12.613
$N = 10000$	$M = N/4$	6.3880	52.257	51.680	27.864
	$M = N/2$	5.9462	38.256	53.707	30.924

From Tables 4 and 5, although the errors of $\|Ax - b\|$ and $\|x_{\text{orig}} - x\|_{\infty}$ resulted from *CSMP* and *SP* are basically equal to zero, the others stemmed from *GSPA* and *NIHT* are approximately close to zero as well, and thus there is no big distinction of recovered effects among those four methods. However, one can not be difficult to find that in Table 6 the average CPU time cost by *GSPA* is much lower than those spent by three other methods, which means under such circumstance our proposed approach run extremely fast. For instance when $N = 10000$ with $M = N/2$ ($M = N/4$) and $s = 5\%N$, the CPU time only need 5.9462(6.3880) seconds via *GSPA*, while 38.256(52.257), 53.707(51.680), 30.924(27.864) seconds yielded by *NIHT*, *CSMP* and *SP* respectively.

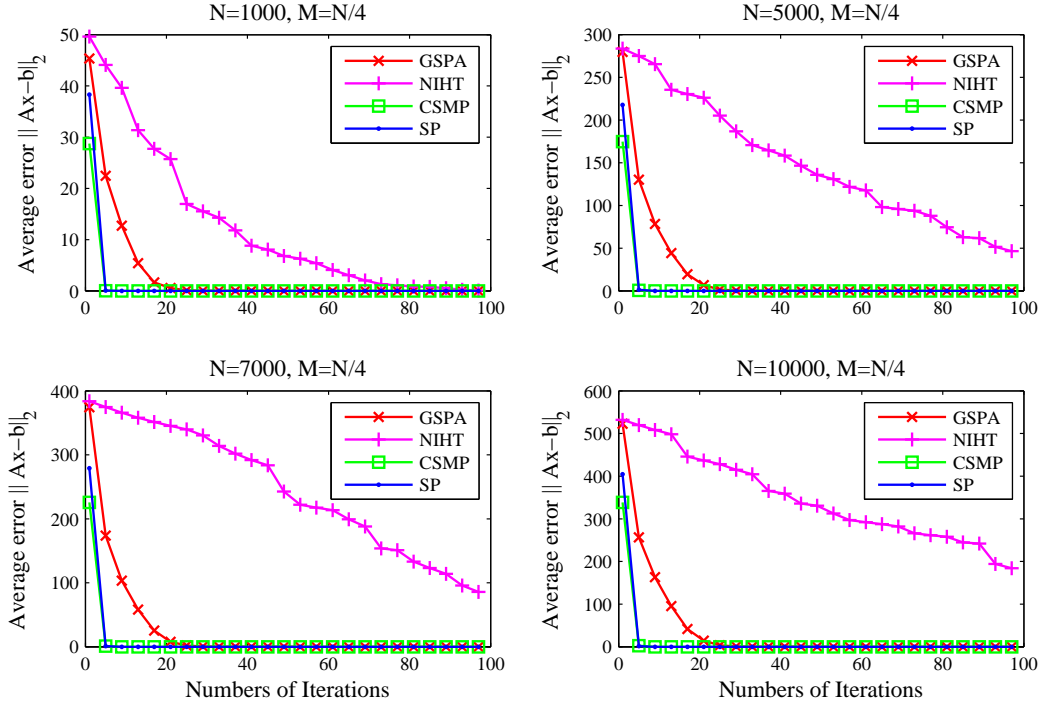


Figure 2: Average prediction error $\|Ax - b\|$ for each iteration with $s = 5\%N$ over 40 simulations.

We then run 40 simulations to count the average error $\|Ax - b\|$ for each iteration. The first 100 iterations will be taken into account to observe the average descent rate of the error $\|Ax - b\|$ from four algorithms. Seeing Figure 2, *NIHT* requires the far over 100 iterative times to make the error $\|Ax - b\|$ decline to a desirable level. Compared with that, *GSPA* almost need 20 iterative times for any dimensions N to reach the lowest level. Even though times of iterations (nearly 7 times for each N) demanded by *CSMP* and *SP* are smallest among these four algorithms, it also indicates that each iteration must cost a relatively long time based on the CPU time in Table 6.

Recovery with Noise: For the sake of clarity of illustrating the robust of these algorithms, we

sequently simulate the recovery with noisy case. Corresponding MATLAB codes are modified to:

$$\begin{aligned} x_{\text{orig}} &= \text{zeros}(N, 1), \quad y = \text{randperm}(N), \quad x_{\text{orig}}(y(1:s)) = \text{randn}(s, 1), \\ A &= \text{randn}(M, N), \quad b = A * x_{\text{orig}} + \sigma_0 * \text{randn}(M, 1), \end{aligned}$$

where the noise obeys to the normal distribution with zero expectation and σ_0^2 (taken as $\sigma_0 = 0.01$ for simplicity) variance. Specific figures produced by these four approaches when $M = N/4$ and $k = 5\%N$ are recorded in Table 7, where "–" denotes the invalid computation. The most significant property of the data in the table is the recovered effects ($\|Ax - b\|$ or $\|x_{\text{orig}} - x\|_{\infty}$) of *GSPA*, *NIHT*, *CSMP* and *SP* are almost nondistinctive. In other words, with noise disturbing, *CSMP* and *SP* no longer perform as well as that in absence of noise. Particularly, when the sample size $N \geq 5000$, *CSMP* behaves extremely worse so that it is impossibility implementary in the high dimensional real applications. What makes the results stunning in Table 7 is that the average CPU time of *GSPA* is far of smallness, comparing with time spent by *NIHT*, *CSMP* and *SP*, which indicates these three methods are not appealing when the affine equations are interfered by some noise, even though the noise is quite minute.

Table 7: Average results over 40 simulations with $M = N/4$, $s = 5\%N$ and noise.

	N	<i>GSPA</i>	<i>NIHT</i>	<i>CSMP</i>	<i>SP</i>
$\ Ax - b\ $	1000	0.1376	0.1376	0.1723	0.1376
	3000	0.2505	0.2505	0.3056	0.2505
	5000	0.3216	0.3216	--	0.3216
	7000	0.3718	0.3716	--	0.3725
	10000	0.4478	0.4478	--	0.4487
$\ x_{\text{orig}} - x\ _{\infty}$	1000	0.0022	0.0022	0.0032	0.0022
	3000	0.0012	0.0012	0.0020	0.0012
	5000	0.0010	0.0010	--	0.0010
	7000	0.0007	0.0008	--	0.0011
	10000	0.0008	0.0008	--	0.0009
CPU time	1000	0.0812	0.3226	116.87	0.1859
	3000	0.5797	3.9317	1416.1	1.1631
	5000	1.6221	9.6857	--	4.9076
	7000	3.2252	25.306	--	11.556
	10000	6.6369	38.440	--	28.429

In Figure 3, for the comparison between *GSPA* and *NIHT*, one can check that the average prediction error $\|Ax - b\|$ begins to close to zero when *GSPA* iterates nearly 20 steps, which is smaller than that *NIHT* does. When it comes to compare *GSPA* and *CSMP*, we reduce the sample size due to the time complexity of *CSMP* (see Table 7). Although at the beginning the error of *CSMP* descends dramatically (here 1-10 iterative times has not been plotted in middle of Figure 3), then it almost sta-

bilizes at a small error and does not decline to zero again. By contrast, the error from *GSPA* always drops until to zero. In terms of comparison between *GSPA* and *SP*, one can observe the iterative times (approximately 7 times) for *SP* to reach the bottom are relatively small, whilst the error from *GSPA* requires nearly 10 (30) times when $M = N/2$ ($M = N/4$) to reach the lowest point. However, meticulous readers are not difficult to find that based on the CPU time in Table 7, the time for each iteration of *SP* must cost longer than *GSPA*.

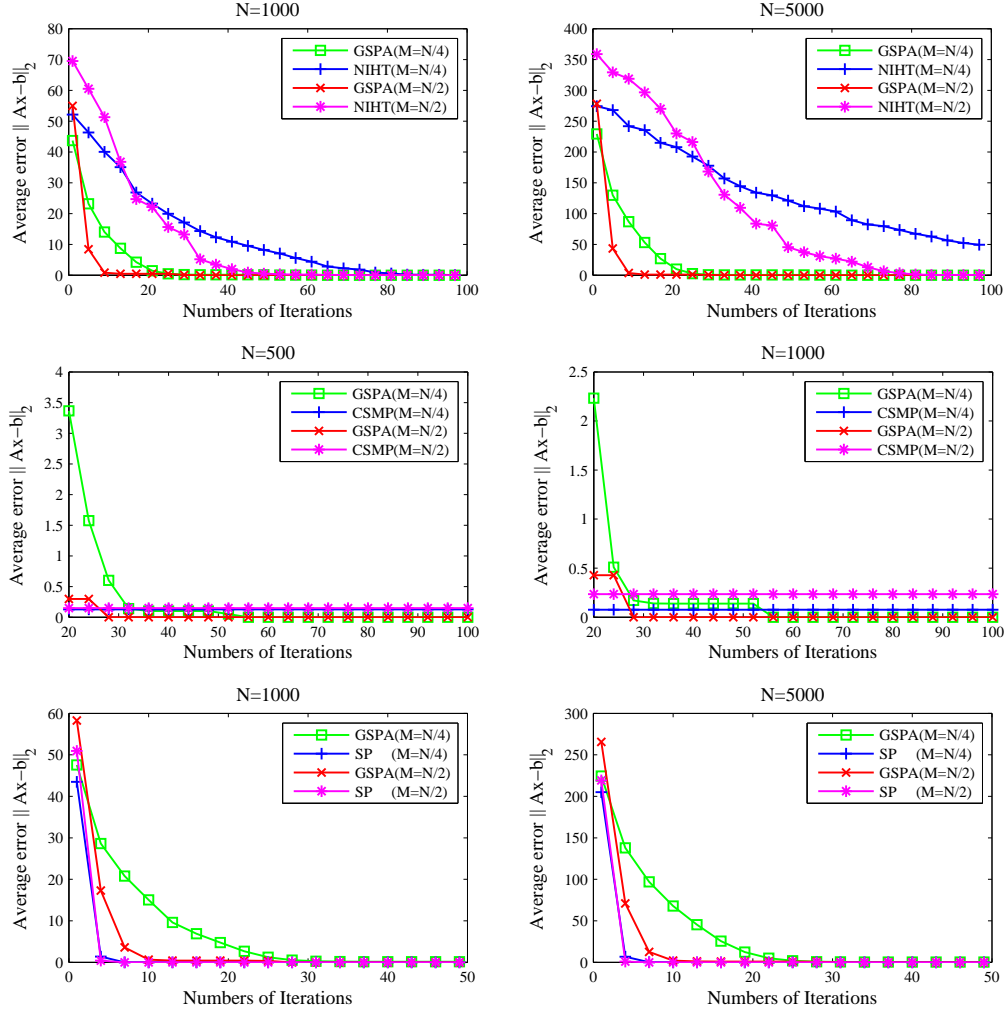


Figure 3: Average error $\|Ax - b\|_2$ for each iteration with $s = 5\%N$ over 40 simulations with noise.

Since the fact that *CSMP* and *SP* would perform worse under the relatively larger sparsity of x_{orig} , we consider the sparsity s as $s = 1\%N$ under the noisy case. The information in Table 8 shows that when the sparsity s of x_{orig} is far less than N ($s = 1\%N$), *CSMP* and *SP* will perform as robustly as *GSPA* and *NIHT* do, because the corresponding results in Table 8 of these four methods basically tend to be similar.

Table 8: Average results over 40 simulations with $M = N/4$, $s = 1\%N$ and noise.

	N	$GSPA$	$NIHT$	$CSMP$	SP
$\ Ax - b\ $	1000	0.1549	0.1548	0.1565	0.1549
	3000	0.2752	0.2752	0.2782	0.2752
	5000	0.3418	0.3418	0.3455	0.3418
	7000	0.4073	0.4073	0.4129	0.4073
	10000	0.4889	0.4889	0.4942	0.4889
$\ x_{\text{orig}} - x\ _{\infty}$	1000	0.0011	0.0011	0.0015	0.0012
	3000	0.0007	0.0007	0.0010	0.0007
	5000	0.0007	0.0007	0.0009	0.0007
	7000	0.0007	0.0007	0.0008	0.0007
	10000	0.0006	0.0006	0.0007	0.0006
CPU time	1000	0.0232	0.0543	0.0455	0.0134
	3000	0.1368	0.3334	0.1559	0.0739
	5000	0.3383	1.1299	0.5412	0.1848
	7000	0.6440	1.9878	1.6966	0.5325
	10000	1.4000	4.5121	3.3096	1.1757

5.3 Comments

From these two comparisons: comparison of N_{NIHT} and N_{GSPA} and comparison of $GSPA$, $NIHT$, $CSMP$ and SP , some comments can be concluded.

- There is no essential distinction between our N_{GSPA} and $GSPA$, because the projection on a nonnegative cone does not obstruct the computational time and recovered effects. From experiments and analysis above, the proposed method $GSPA$ performs very steadily, and thus does not be overly relied on the sample size M and N . It also runs relatively well for some different sparsity s of x_{orig} . In addition, regardless of the exact recovery and case with noise, $GSPA$ unravels its good robustness. Importantly, $GSPA$ is the most fast of all these four approaches;
- For exact recovery, $NIHT$, $CSMP$ and SP all proceed a good performance, particularly the two latter approaches enable the recovery to be exceptionally exact (i.e., making the error $\|Ax - b\|$ and $\|x_{\text{orig}} - x\|_{\infty}$ extremely equal to zero), but the recovered effect of $NIHT$ quite depends on the sample size M and N . When referring to the recovery with noise, the recovered effects from $CSMP$ and SP are no longer better than $GSPA$ and $NIHT$, particularly the performance of $CSMP$ which excessively relies on the sparsity s are becoming much worse. Moreover, the CPU time generated by these three methods is all far higher than that needed by $GSPA$, which implies in high dimensional recovery they would not be appealing.

6 Conclusion

In this paper, we have established the first and second order optimality conditions for model (2) and (3), proposed a gradient support projection algorithm for AFP_{SN} , and shown that the new algorithm has elegant convergence and exceptional performance. In the future, we will develop this algorithm for solving splitting feasibility problem (by Censor in [12]) with sparsity and other complex constraints.

Acknowledgements

We are grateful to Dr. Caihua Chen in Nanjing University for his helpful advice. The work was supported in part by the National Basic Research Program of China (2010CB732501), and the National Natural Science Foundation of China (11171018, 71271021).

References

- [1] Attouch H, Bolte J and Svaiter B F 2013 Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods *Mathematical Programming* **137** 91-129
- [2] Bahmani S, Raj B and Boufounos P T 2013 Greedy sparsity-constrained optimization *Journal of Machine Learning Research* **14** 807-41
- [3] Baraniuk R, Cevher V, Duarte M and Hegde C 2010 Model-based compressive sensing *IEEE Transactions on Information Theory* **56** 1982-2001
- [4] Beck A and Eldar Y 2013 Sparsity constrained nonlinear optimization: optimality conditions and algorithms *SIAM Journal on Optimization* **23** 1480-509
- [5] Blumensath T and Davies M 2006 Sparse and shift-invariant representations of music, *IEEE Transactions on Audio, Speech and Language Processing* **14** 50-57
- [6] Blumensath T and Davies M 2008 Iterative thresholding for sparse approximations *Journal of Fourier Analysis and Applications* **14** 626-54
- [7] Blumensath T and Daviex M E 2010 Normalized iterative hard thresholding: Guaranteed stability and performance *IEEE Journal of Selected Topic in Signal Processing* **4** 298-309
- [8] Bruckstein A M, Donoho D L and Elad M 2009 From sparse solutions of systems of equations to sparse modeling of signals and images *SIAM Review* **51** 34-81

- [9] Candés E J and Tao T 2005 Decoding by linear programming *IEEE Transactions on Information Theory* **51** 4203-15
- [10] Calamai P H and Moré J J 1987 Projection gradient methods for linearly constrained problems *Mathematical Programming* **39** 93-116
- [11] Cartis C and Thompson A 2013 A new and improved quantitative recovery analysis for iterative hard thresholding algorithms in compressed sensing arXiv:1309.5406
- [12] Censor Y and Elfving T 1994 A multiprojection algorithm using Bregman projections in a product space *Numerical Algorithms* **8** 221-39
- [13] Dai W and Milenkovic O 2009 Subspace pursuit for compressive sensing signal reconstruction *IEEE Transactions on Information Theory* **55** 2230-49
- [14] Davis G, Mallat S and Avellaneda M 1997 Adaptive greedy approximations *Constructive Approximation* **13** 57-98
- [15] Donoho D L and Tanner J 2005 Sparse nonnegative solutions of underdetermined linear equations by linear programming *Proceedings of the National Academy of Sciences of the United States of America* **102** 9446-51
- [16] Foucart S 2011 Hard thresholding pursuit: an algorithm for compressive sensing *SIAM Journal on Numerical Analysis* **49** 2543-63
- [17] He R, Zheng W, Hu B and Kong X 2011 Nonnegative sparse coding for discriminative semi-supervised learning in *Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR)* 2849-56
- [18] Mallat S and Zhang Z 1993 Matching pursuits with time-frequency dictionaries *IEEE Transactions on Signal Processing* **41** 3397-415
- [19] Needell D and Tropp J A 2009 CoSaMP: Iterative signal recovery from incomplete and inaccurate samples *Applied and Computational harmonic Analysis* **26** 301-32
- [20] Rockafellar R T and Wets R J 1998 *Variational analysis* Springer, Berlin